AUTOMATIC TRANSCRIPTION OF ENGLISH BROADCAST NEWS

Peter Beyerlein, Xavier Aubert, Reinhold Haeb-Umbach, Dietrich Klakow, Meinhard Ullrich, Andreas Wendemuth and Patricia Wilcox

Philips Research Laboratories
Weisshausstr. 2, D-52066 Aachen, Germany, beyerlei@pfa.research.philips.com

ABSTRACT

In this paper the Philips Broadcast News transcription system is described. The Broadcast News task aims at the recognition of "found" speech in radio and television broadcasts without any additional side information (e.g. speaking style, background conditions). The system was derived from the Philips continuous mixture density crossword HMM system, using MFCC features and Laplacian densities. A segmentation was performed to obtain sentence-like partitions of the broadcasts. Using data-driven clustering, the obtained segments were grouped into clusters with similar acoustic conditions for adaptation purposes. Gender independent word-internal and crossword triphone models were trained on 70 hours of the HUB4 training data. No focus condition specific training was applied. Channel and speaker normalization was done by mean and variance normalization as well as VTN and MLLR. The transcription was produced by an adaptive multiple pass decoder starting with phrase-bigram decoding using word-internal triphones and finishing with a phrase-trigram decoding using MLLR-adapted crossword models.

1. INTRODUCTION

Past speech recognition research has mainly focused on the decoding of good quality speech in clean environments. The focus has recently shifted to speech "found" in the "real world". One of the data sources of real-world speech are audio records from radio and television broadcast news. The data in these broadcasts have basically the following additional characteristics:

- Unknown sentence boundaries (if the parts of speech to be recognized can be called sentences).
- Diverse and rapidly changing acoustic environment. Typical degradations of the quality of the speech signal are introduced by background music, noise, interfering speakers as well as by changes between studio and telephone channels. On top of this, regional dialects or accents of non-native speakers have to be considered.
- Real-life speaking styles (spontaneous speech) as well as speaker turns. Speaking styles range from carefully read speech to free and spontaneous conversation.
- Natural language. Difficulties arise from unpredictable changes of topics of the broadcast news as well as from unpredictable reactions in free conversation.

Thus the challenges of processing broadcast news data may be summarized as follows:

- How to break the broadcast signal stream with a duration of up to 180 minutes down to "sentences" (called segments) with consistent channel, background and speaker characteristics? In addition, the segment boundaries should ideally be consistent with the linguistic boundaries of the spoken stream of words.
- How to detect, reduce and learn the variation of the signal caused by rapid changes of the channel, the background and the speaker?
- How to reduce the variation of the signal caused by changes of the speaking style?
- How to predict word sequences in conversational English language?

This paper summarizes our approach to deal with the new problems and describes the system we developed in 1997.

2. SEGMENTATION

2.1. Segmentation and Classification

The automatic segmentation provided by NIST was used for a first chopping of the whole (three-hours) audio stream. The resulting 769 segments were subsequently processed by a phone decoder [1] to produce a refined set of segments automatically labeled in terms of male speech, female speech and non-speech. The context-independent phoneme models of the phone decoder were gender-dependently trained on the Broadcast News (BN) training data augmented with non-speech data (20 minutes of music and noise). The phone decoder was a one-pass Viterbi beam-search decoder that evaluates male and female models in parallel, guided by a bigram phoneme model. The output thereof was post-processed to produce homogeneous speech intervals tagged as male or female as well as non-speech segments. Table 1 shows the resulting distribution of the segments.

#segments	all	female	male
small band	271	66	205
wide band	1160	417	743
non-speech	227	-	-
all	1658	483	948

Table 1: *Segmenter output*

The non-speech segments represent a total of 8.01 minutes of data that were eliminated from subsequent decoding stages. Speech was also classified in terms of large or small bandwidth using the F0/F2 labels of the NIST-provided segments. However, this classification was only used in the segment clustering and not in the decoding.

2.2. Segment Clustering

The speech segments were clustered using an agglomerative bottom-up technique based on the relative cross-entropy measure. In this evaluation, we used the between-segment distance implemented in the CMU segmenter [2], namely, the Kullback-Leibler distance metric, augmented with an additive term to favore the merging of adjacent segments. Agglomerative clustering is performed in two stages, a nearest-neighbor search controlled by a small distance threshold followed by a search subjected to a larger threshold for satisfying a minimum duration constraint. This procedure was applied separately on male and female segments for both large and small speech bandwidths. This provided 106 clusters of segments with an average duration of 100 seconds, the minimum length being set to 25 seconds.

2.3. UE versus PE

We compared the generated automatic segmentation (UE) with the manual segmentation provided by NIST (PE) on the HUB4'96 evaluation task. In Tab. 2 error rates are reported for the first bigram decoding pass (I) and the final trigram crossword decoding pass (VI). A more detailed description of the decoding passes can be found in section 7. The word error rate increased on average by about 5% using the generated automatic segmentation.

		CNN	CSP	NPR W.	NPR M.P.
	file1-4	file1	file2	file3	file4
PE, I	35.0	36.7	33.4	39.7	30.3
UE, I	36.3	37.1	35.3	40.4	32.4
PE, VI	28.7	29.0	26.4	34.2	25.6
UE, VI	30.2	30.0	29.1	35.1	26.8

Table 2: *Word error rates in % on HUB4'96 eval. set for different segmentation scenarios.*

3. FEATURE EXTRACTION

This section gives an overview of the acoustic front end used for the HUB4'97 evaluation. For more detailed investigations related to this front end see [13].
In the acoustic front end, mel-frequency cepstral coefficients are computed. A feature vector consists of 15 static features, 15 linear regression delta features, the frame energy and its first- and second-order derivatives, resulting in a 33-component feature vector.

3.1. LDA

Linear discriminant analysis is used to reduce the scatter of the feature vectors within the classes compared to the overall scatter. Three consecutive feature vectors are adjoined to a 99-component vector to which a linear discriminant transformation is applied. The gender-independent LDA matrix was estimated using the BN training data. The 35 vector components with the largest eigenvalues were retained to form the final feature vector. It is interesting to note that the LDA trained on the BN training data performs only slightly better than the LDA trained on the WSJ training data: on the HUB4'96 dev. set the error rate dropped from 36.9% to 36.2%. The performance improvement obtained by an LDA matrix trained on HUB4 data was not big, however consistent over most focus conditions.

3.2. VTN

Vocal tract normalization (VTN) was applied in training and recognition. The intention was to reduce the influence of differences in the vocal tract length between speakers on the computed feature vector. A linear warping of the frequency axis was realized by suitably shifting the center frequencies of the mel-frequency filter bank. For the warping factor selection we adopted a maximum-likelihood approach, see e.g. [11]. In recognition, the hypothesized transcription required by VTN was obtained from a first bigram decoding without VTN. From the results of Table 3 we concluded that VTN provides a means to overcome the need for gender-dependent (GD) models. We therefore decided to use a GI setup with VTN in training and recognition for the HUB4'97 evaluation.

Setup			WER
GD	no	VTN	36.3
GD	+	VTN in recognition	35.6
GI	+	VTN in training & recognition	35.4

Table 3: *Effects of VTN in training and recognition on the word error rate for gender-dependent (GD) and gender-independent (GI) models. HUB4'96 evaluation test set, UE, bigram decoding*

3.3. Feature Normalization

The same signal analysis was applied to all data, i.e. there is no special treatment for the small-bandwidth (F2) data. The cepstral features were normalized for each segment by cepstral mean subtraction and by unit variance normalization. On the HUB4'96 development data a reduction of word error rate of 3% relative due to variance normalization was observed.

4. LEXICAL MODELING

The recognition lexicon is an extension of the Philips NAB 64k lexicon and consists of 74,000 entries. New words were transcribed by an automatic transcription system [7] and some of these were corrected manually. Phrase models were applied both in training and recognition (a phrase is a word sequence which frequently occurs in the training data). Each phrase was included into the lexicon and into the language model as a single entry. Phrases are a simple means of modeling long-span acoustic and language context. We modeled typical variations in speaking style and coarticulation of the most frequent word sequences by adding pronunciation variants to the phrases in the pronunciation lexicon. The HUB4 training and recognition lexica were augmented with the 330 most frequent phrases found in the BN training data. The 10 most frequent phrases found in the BN training data are: *in_the, of_the, on_the, to_the, and_the, you_know, for_the, to_be, I_think, that_the*. Each word in the lexicon has on the average 1.15 pronunciation alternatives.

4.1. Motivation to use Phrases

We investigated the effect of phrase modeling on the Wallstreet-Journal task. The obtained perplexities are presented in Tab. 4. The vocabulary size is 5 K words. For 226 phrases added to the vocabulary the trigram perplexity ($M = 3$) is reduced by 7.2%. (Note that all perplexities reported are normalized to words.) For the fourgram the improvement is still 3.0%. However, the main effect of phrases is not the reduction in perplexity but the reduction in word error rate due to an improved pronunciation modeling of frequent word sequences. Word error rates for cross-word decoding for the WSJ task are reported in Tab. 5. It is obvious that there is a reduction in word error rate larger than expected from the perplexity reductions. Also, this reduction in word error rate persists for bigram, trigram and fourgram decoding.

# Phrases	0	226
M=1	738.0	562.6
M=2	113.0	100.0
M=3	60.8	56.4
M=4	52.6	51.0
M=5	50.4	50.3

Table 4: *Perplexities for the WSJ task.*

Model	M=2	M=3	M=4
5 K words	8.2%	7.0%	6.9%
+ 226 phrases	7.7%	6.1%	6.0%

Table 5: *Word error rate on si_dt_05'92, si_et_05'92, si_dt_05'93, si_et_h2'93 for bigram, trigram and fourgram language models showing the influence of phrase modeling.*

5. LANGUAGE MODELING

5.1. Language Modeling Using Phrases

The training corpus consists of 140 million words of transcribed broadcast news. The 1996 development data were used as test set. Perplexities are shown in Tab. 6. Due to the phrases the perplexity of the bigram model was reduced by 8.4% and the perplexity of the trigram model was reduced by 4.1%. As a result of the phrase modeling the average context length of the trigram language model was 3.5 words.

	M=1	M=2	M=3
64 K words	1026.4	257.1	180.0
+ 330 phrases	841.2	235.4	172.7

Table 6: *Perplexities for the HUB4'96 development set.*

5.2. Language Model Adaptation

For the HUB4'97 evaluation, language model adaptation techniques [8] were used to adapt and combine statistics gained from the following corpora: Broadcast News (BN), North American Business News (NAB) and the transcripts of the acoustic training material (TAT). The transcripts of the acoustic training data were mainly used as a cross-validation set. On each corpus a language model was estimated. The FMA technique [5],[8] was used for an adaptation of the NAB language model to the HUB4 task . Adaptive linear interpolation was the preferred method to combine the three models.

5.3. Results on the Evaluation Data

This section briefly summarizes key figures for the 1997 evaluation data. 831 words from TAT and from the speaker data base were supplemented to the vocabulary. The chosen vocabulary resulted in an out-of-vocabulary(OOV) rate of 0.48% on the 1997 evaluation data. Tab. 7 gives the perplexities for the language models described above on the 1997 evaluation data. For the bigram, addition of phrases gave a 9% improvement and the adaptive combination another 9%.

Model	Perplexity
M=2 BN without phrases	236.3
M=2 BN	215.7
M=2 BN + TAT + NAB adap. (*)	195.9
M=3 BN	149.9
M=3 BN + TAT (*)	146.6

Table 7: *Perplexities on HUB4'97, (*) = used in evaluation*

We observed that the additional gain when going from a trigram to a fourgram was small. Thus we decided to use the 3.5-gram (phrase 3-gram) language model only.

6. ACOUSTIC TRAINING

6.1. Acoustic Modeling

In the acoustic modeling we employed continuous mixtures of Laplacian densities with a single, globally pooled deviation vector. Decision tree clustering was applied to continuous single Laplacian densities for within-word and crossword triphones [3]. We used the question set proposed by Odell [4]. As goodness-of-split criterion we took the maximum likelihood approach with a few approximations which eventually led to a simple minimum mean distance criterion. We obtained 8k clusters and 340k densities ($11.9 \cdot 10^6$ parameters) for 33k gender independent within-word triphone states and 10k clusters and 420k densities ($14.7 \cdot 10^6$ parameters) for 81k gender independent crossword triphone states.

6.2. Training Strategy

We trained two gender-independent model sets on the BN training data, one for within-word and one for crossword decoding. There were no focus-specific model sets. Before taking this decision we compared several strategies on the HUB4'96 dev. set:

1. Gender-dependent training on the wsj0+1 training data and subsequent supervised adaptation specifically on each of the HUB4 focus conditions.

2. Gender-dependent training of a separate model set for each of the HUB4 focus conditions.

3. Gender-dependent training of one model set on all available HUB4 data.

We observed that all three strategies performed similarly (see Table 8), the simplest strategy (scenario 3) performed best.
Thus we decided to use only focus condition independent model sets (scenario 3). Finally using VTN in training and recognition gender independent word-internal and crossword triphone models were trained on 70 hours of the HUB4 training data.

Scenario	HUB4'96 dev. set
1	41.9
2	42.4
3	38.6

Table 8: *Word error rates in % on HUB4'96 dev. set (male speakers only) for different training scenarios. Bigram language model, gender-dependent setup, no adaptation in recognition.*

6.3. Unsupervised MLLR Adaptation

Unsupervised MLLR adaptation of the mean vectors was applied to clusters of segments using the Least Mean Square Approximation [12]. For information on the clustering procedure, see section 2. The regression classes are based on phonetic knowledge and are dynamically defined using a tree organization. The amount of adaptation speech determines both the number of active regression classes and the structure of the MLLR transformation matrices. In light of the presumably high error rate we adopted a conservative approach and used more than one MLLR transformation matrix only for classes with more than 10,000 observations. We used a single block-diagonal or purely diagonal matrix if the number of observations was below 1000 and 200, respectively. Note that MLLR adaptation was applied to both the within-word model set and the crossword model set. Table 11 presents the results for adaptation of the mean vectors of the within-word models. It can be seen that the error rate improvement due to VTN and MLLR was about 8% on the HUB4'97 evaluation data.

7. DECODING RESULTS

Decoding was done in a number of stages [3]:

I a Using the hypothesized word sequence from a preliminary bigram decoding, a linear frequency warping factor was estimated for each segment and the features were warped accordingly (VTN). For the decoding gender-dependent within-word triphone HMM's were used, which were trained without VTN.

I b A time-synchronous Viterbi bigram decoding was carried out, producing a bigram lattice as its output. For the decoding gender-independent within-word triphone HMM's were used, which were trained on VTN features.

II Using the hypothesized word sequence from the bigram decoding (I), VTN was applied. The obtained features were used in the further processing stages.

III MLLR adaptation was applied to the respective clusters of segments. A new bigram decoding was then done on the previously generated lattice employing the adapted models and resulting in a new output lattice.

IV Trigram rescoring was carried out on that lattice and the lattice was pruned employing the N-best paradigm. In order to obtain sufficient variability even for fairly small N, the original segments were further subdivided into shorter "subsegments". The number K of subsegments depends on the length of the segment and the number of pauses within the segment. N-best sentence hypotheses were generated for each of the subsegments, and then the subsegments were again concatenated to form a full segment. Thus we obtained the N^K-best sentences instead of the N-best sentences for each of the segments. Using this technique, the lattice error rate after the N-best purging was reduced by a factor of two to three on the development data.

V The resulting purged lattice was input to the trigram cross-word decoder [3].

VI Batch unsupervised MLLR adaptation was carried out on the cross-word models, with the hypothesized transcription obtained from the last pass. For MLLR adaptation of the crossword models the same parameters were used as for the adaptation of the within-word triphone models. A new crossword trigram decoding was conducted using the adapted models, resulting in the final system output.

		broadcast			
		CNN	CSP	NPR W.	NPR M.P.
	file1-4	file1	file2	file3	file4
PE,I	35.0	36.7	33.4	39.7	30.3
PE,III	32.0	32.7	30.5	36.6	28.5
PE,IV	30.2	30.6	28.1	36.2	26.3
PE,VI	28.7	29.0	26.4	34.2	25.6

Table 9: *Word error rates in % on HUB4'96 PE eval. set for various decoding passes*

		broadcast			
		CNN	CSP	NPR W.	NPR M.P.
	file1-4	file1	file2	file3	file4
UE,I	36.3	37.1	35.3	40.4	32.4
UE,VI	30.2	30.0	29.1	35.1	26.8

Table 10: *Word error rates in % on HUB4'96 UE eval. set for various decoding passes*

	HUB4-E 97
UE,I	29.0
UE,III	26.7
UE,IV	24.8
UE,VI	23.1 (23.5)

Table 11: *Word error rates in % on HUB4'97 English evaluation set for various decoding passes*

Tables 9, 10, 11 summarize the performance of the Philips system for the HUB4'96 and the HUB4'97 evaluation tasks using the original release (Nov. 1997) of the HUB4 scoring rules. The official result of the HUB4'97 English evaluation system was 23.5 % word error rate. After correcting a bug in the postprocessing of the final system output we obtain 23.1% word error rate.

References

1. F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, J. Makhoul, "Advances in Transcription of Broadcast News", in Proc. Eurospeech pp.927-930, Rhodes, Greece, Sep. 1997.
2. M. Siegler, U. Jain, B. Raj, R.M. Stern, "Automatic Segmentation and Clustering of Broadcast News Audio", in Proc. of the DARPA Speech Recognition Workshop, pp. 97-99, Westfields, Chantilly, Virginia, Feb. 2-5, 1997.
3. P. Beyerlein, M. Ullrich, P. Wilcox, "Modelling and Decoding of Crossword Context Dependent Phones in the

Philips Large Vocabulary Continuous Speech Recognition System", in Proc. EUROSPEECH, pp. 1163-1166, Rhodes, Greece, Sep. 1997.

4. J. J. Odell: "The Use of Context in Large Vocabulary Speech Recognition", Ph.D. thesis, University of Cambridge 1995, England.

5. R. Kneser, J. Peters, and D. Klakow. "Language model adaptation using dynamic marginals", in: Proc. EUROSPEECH, pp. 1971–1974, Rhodes, Greece, Sep. 1997

6. J.L. Gauvain, G. Adda. L. Lamel, and M. Adda-Decker: "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System", *DARPA Speech Recognition Workshop*, pp. 56, 1997.

7. S. Besling: "A Statistical System for Grapheme–to–Phoneme Conversion", *Proc. Tenth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research: Reflections on the Future of Text*, pp. 5, 1994.

8. D. Klakow, X. Aubert, P. Beyerlein, R. Haeb-Umbach, M. Ullrich, A. Wendemuth, P. Wilcox, "Language Model Investigations Related To Broadcast News", *elsewhere in these proceedings*

9. R. Schwartz, H. Jin, F. Kubala, and S. Matsoukas, "Modeling those F-Conditions - or not", in Proc. DARPA Speech Recognition Workshop, pp 115-119, Chantilly, VA, Feb. 1997.

10. R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", in Proc. ICASSP'92, pp I13-I16, San Francisco, CA, March 1992.

11. L. Lee, R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP* Vol. 1, pp. 353-356, Atlanta, GA, May 1996.

12. E. Thelen, X. Aubert and P. Beyerlein, "Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", in Proc. ICASSP, pp. 1035-1038, Munich, Germany, April 1997.

13. R. Haeb-Umbach, X. Aubert, P. Beyerlein, D. Klakow, M. Ullrich, A. Wendemuth, P. Wilcox, "Acoustic Modeling in the Philips HUB4-System", elsewhere in these Proceedings.